Terry Lyu 435091 wustl key terrylu

Q1.

Part a

there are four files created, indicating that there are four reducers used. the first few lines are

```
SEQ!org.apache.hadoop.io.LongWritableorg.apache.hadoop.io.Text)org.apache.hadoop.io.compr
ess.SnappyCodecw
```

it indicates that the sequence file has been created, and we use snappy to compress the file.

The uncompressed file size is 582639761, the compressed file size is 136557523, so the compression ratio is (582639761/136557523) = 4.2666

part c

after uncompressing, we can actually see the weblog files instead of the unreadble compressed version of sequence files.

part d

```
//        FileOutputFormat.setCompressOutput(job,true);
      // FileOutputFormat.setOutputCompressorClass(job,SnappyCodec.class);

    //    SequenceFileOutputFormat.setOutputCompressionType(job, CompressionType.BLOCK);
```

```
hadoop jar param.jar \ -Dmapred.output.compress=true \

-Dmapred.output.compression.codec=org.apache.hadoop.io.compress.SnappyCodec \
weblog paramcompressout
```

Q2.

1. The type is KeyValueTextInput ,

the code to retrieve filename is

```
FileSplit fsplit = (FileSplit) context.getInputSplit();
            Path path = fsplit.getPath();
            String fileName = path.getName();
```

2. see code submitted

3.mapper output is

```
Have Hamlet@282
heaven Hamlet@282
and Hamlet@282
earth Hamlet@282
together Hamlet@282
There Hamlet@133
are Hamlet@133
more Hamlet@133
things Hamlet@133
in Hamlet@133
heaven Hamlet@133
and Hamlet@133
earth Hamlet@133
```

Q3

Different place for conjunctions or phrases could have the same meaning in modern English but could get different result from word cooccurence .

Ex: though now is not a good time v.s. now is not a good time though

we should use Filesplit and getinputsplit to get the record split between two blocks to make sure that sentences potentially broken apart and stored in two different blocks are processed correctly.